Cell PRESS
Open ACCESS

EXTENDED EXPERIMENTAL PROCEDURES

Protein Expression and Purification

Full-length *CBF1*and *TYE7* open reading frames cloned into the Gateway pDEST15 (N-terminal GST-tag) expression vectors (Invitrogen) were obtained from (Zhu et al., 2009).

GST-Cbf1 and GST-Tye7 were overexpressed in *E. coli* BL21 (DE3) cells (New England BioLabs) and purified by FPLC (AKTAprime plus) using 1 ml GSTrap™ FF affinity columns (GE Healthcare). Samples were concentrated by centrifugation using Amicon Ultra filters (Millipore), glycerol was added to a final concentration of 10%, and proteins were stored at −80°C until further use. Western blots were performed for each protein to assess quality and to approximate protein concentration by visual inspection relative to a dilution series of recombinant GST standard (Sigma), as described previously (Zhu et al., 2009).

gcPBM Design

We designed a custom DNA microarray (Agilent Technologies, Inc.; AMADID #029393) containing putative Cbf1 and Tye7 DNA binding sites within ~44,000 probes. The probes are 60 bp long and have the general form $CAT(N)_{30}TTCGTCTTGATT$ CGCTTGACGCTGCTG, where the last 24 nucleotides are complementary to a common primer used to double-strand the oligonucleotide array by primer extension (Berger et al., 2006) and $(N)_{30}$ corresponds to 30-bp sequences from the *S. cerevisiae* genome. The constant 3-bp sequences immediately flanking the genomic sequences (CAT and TTC) were selected so that they do not inadvertently create putative Cbf1 or Tye7 binding sites outside of the selected genomic regions. We chose to represent three categories of 30-bp genomic sequences on our gcPBM: 1) "ChIP-chip bound" probes, 2) "ChIP-chip unbound" probes, and 3) negative control probes. To generate the "ChIP-chip bound" probes, we first identified the genomic regions bound in vivo by Cbf1 or Tye7 (*i.e.*, sequences with ChIP-chip $p < 0.005$ in rich medium (YPD) (Harbison et al., 2004)), and then searched those sequences for putative TF binding sites, defined as 10-bp sites that contain at least two consecutive 8-mers with PBM E-score > 0.35 according to the Cbf1 or Tye7 universal PBM data reported by (Zhu et al., 2009). These 10-bp sites were aligned to 10-bp long Cbf1 and Tye7 DNA binding site motifs (derived from the DNA motifs of (Zhu et al., 2009)) that contained the E-box site and 2-bp flanking regions; this ensured that all putative binding sites occurred at the same location within the probes on the microarray. Finally, we considered the 30-bp genomic sequences centered at the putative Cbf1 or Tye7 binding sites. The "ChIP-chip unbound" probes were selected similarly to "ChIP-chip bound" probes, except that: (1) they were based on genomic regions with ChIP-chip $p > 0.5$ (*i.e.*, not bound significantly), and (2) we required at least two consecutive 8-mers at a more stringent E-score cutoff of 0.4. The negative control probes were selected from *S. cerevisiae* intergenic regions, with the criterion that the maximum E-score over all 8-mers for Cbf1 or Tye7 in each negative control probe was < 0.3, to ensure that these probes do not contain any putative Cbf1 or Tye7 binding sites. In addition to the three categories of 30-bp genomic sequences (*i.e.*, "ChIP-chip bound" probes, "ChIP-chip unbound" probes, and negative control probes), we also designed probes that contain, within constant flanking regions, all 10-bp sequences that satisfy the following criteria: 1) they occur within the "ChIP-chip bound" probes, and 2) they contain the E-box CACGTG. These probes have the form $CCTAACTACTATA(N)_{10}$ ATAGCTTCGTACA (followed by the priming sequence GTCTTGATTCGCTTGACGCTGCTG), and were specifically designed to compare binding of Cbf1 and Tye7 to putative 10-bp binding sites when these sites do not occur within native genomic flanks (Figure 6C). Each probe is represented at 4 separate spots on the custom array. The reported PBM signal intensity for each probe is the median over the PBM signal intensity values over the 4 replicate spots. To design the "validation" array (Agilent Technologies, Inc.; AMADID #041711), we selected 30-bp genomic sequences from our initial custom array, and introduced 1-bp, 2-bp, 3-bp, or 4-bp mutations at various positions in the original genomic sequences; both the wild-type sequences and the resulting mutant sequences are represented on the validation array.

PBM Experiments and Data analysis

Custom-designed microarrays were synthesized (Agilent Technologies, AMADID #029393 and #041711), converted to double-stranded DNA arrays by primer extension, and used in PBM experiments essentially as described previously (Berger et al., 2006; Berger and Bulyk, 2009). All PBM data reported in this study are from experiments performed either on a fresh slide or a slide that had been stripped exactly once (Berger et al., 2008). Microarray scanning and quantification were performed using ScanArray 5000 (Perkin Elmer) for AMADID #029393, and GenePix 4400A (Molecular Devices) for AMADID #041711. In both cases data normalization was performed using masliner (Microarray LINEar Regression) (Dudley et al., 2002) and the Universal PBM Data analysis Suite (Berger and Bulyk, 2009) as previously described (Berger et al., 2006; Berger and Bulyk, 2009).

SVR Analysis

Support Vector Regression (SVR) was run separately for Cbf1 and Tye7. For each TF, we first selected "ChIP-chip bound" and "ChIP-chip unbound" probes centered at the E-box CACGTG. To ensure that no additional binding sites occur in the regions flanking CACGTG, we selected probes (280 for Cbf1, and 312 for Tye7) for which the maximum E-score over all the 8-mers in the flanks was < 0.3. Next, for each selected sequence we combined the two flanking regions (see Figure 4A) and computed the number of occurrences of each 1-mer, 2-mer and 3-mer in the combined flanks. For example, for a given 30-bp sequence, the feature "6-TTT" will be 0 if TTT does not occur at position 6 in either of the flanks, 1 if TTT occurs at position 6 in one of the flanks, and

2 if TTT occurs at position 6 in both flanks. The positions are numbered starting from the center of the CACGTG core (Figure 4A). We thus obtained a sparse feature matrix for each of the two TFs. As target features for the SVR analyses, we used the natural logarithm of the Cbf1 and Tye7 PBM fluorescence signal intensities. Our goal was to use SVR to train linear models that can predict the PBM log signal intensity for each probe based on the sequence features derived from that probe. We used the ε-SVR algorithm implemented in the libSVM toolkit (Chang and Lin, 2011) for all SVR analyses. We performed a grid search using 10-fold and leave-one-out cross-validation to determine the best values for parameters ε and C: C = 0.005 and ε = 0.1 for Cbf1; C = 0.1 and ε = 0.35 for Tye7 (see below). Using these parameters, we trained the final SVR models using all 280 sequences for Cbf1 and all 312 sequences for Tye7. These models were used to predict the PBM log signal intensities for all the probes on the "validation" array. We also performed an SVR analysis using the 312 sequences selected for Tye7, but shuffling the PBM log signal intensities, to determine the range of R^2 values we would obtain on randomized sets of sequences; the best R^2 on randomized sets of sequences was < 0.1 (Figure S4A).

Parameter Search in the SVR Analyses

The ε-SVR implementation in the libSVM package (Chang and Lin, 2011) was used for all SVR analyses. Analyses were run separately for Cbf1 and Tye7, using different sets of features as input. For each transcription factor (TF) and each set of input features, we selected the best values for parameters C and ε by doing a grid search and using cross-validation to evaluate each set of parameters. We started with 10-fold cross validation and tried a wide range of parameters for C (0.0001, 0.00025, 0.0005, 0.00075, 0.001, 0.0025, 0.005, 0.0075, 0.01, 0.025, 0.05, 0.075, 0.1, 0.25, 0.5, 0.75, 1, 2.5, 5,7 .5, 10, 25, 50, 75, 100) and ε (0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1, 1.05, 1.1, 1.15, 1.2, 1.25, 1.3, 1.35, 1.4, 1.45, 1.5). Next, we narrowed down the parameter range for C (0.001, 0.0025, 0.005, 0.0075, 0.01, 0.025, 0.05, 0.075, 0.1, 0.25, 0.5, 0.75, 1, 2.5, 5) and ε (0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5) and used leave-one-out cross-validation to find the best parameter values within these sets. After obtaining the best parameter values, we ran ε-SVR again with the selected parameters using leave-one-out cross-validation, and this time we stored the predicted gcPBM log signal intensity for every input sequence and we calculated the average weights of all the features in the regression model.

For both Cbf1 and Tye7, we tested SVR models using: 1) 1-mer features, 2) 1-mer and 2-mer features, 3) 1-mer, 2-mer and 3-mer features, and 4) 1-mer, 2-mer, 3-mer and 4-mer features. In a 10-fold cross-validation test, using the 280 Cbf1 sequences (see main text) and the parameters optimized using the grid search described above, we obtained R^2 values of: 0.60, 0.67, 0.74, and 0.71 for the four models, respectively. For the 312 Tye7 sequences, the R^2 values were 0.77, 0.78, 0.88, and 0.86 for the four models, respectively. Thus, the models using 1-mers, 2-mers and 3-mers performed best. The fact that the models using 4-mers did not improve the prediction accuracy is not surprising given that they have a very large number of features compared to training examples, and are thus prone to overfitting the training data.

Design of Validation PBM

To test the accuracy of the regression models trained on sequences from the gcPBM, we introduced mutations at various positions in both the proximal and the distal flanks of the 30-bp gcPBM probes, and generated a "validation" PBM containing three categories of probes: 1) "feature mutations" probes, 2) "flank mutations" probes, and 3) negative control probes.

To generate the "feature mutations" probes, we focused on the sequence features that had large positive or negative weights in the Cbf1 or Tye7 regression models. Since we can only fit a limited number of probes on a custom PBM, we sorted the features in decreasing order of their SVR weights and we selected: 1) proximal features (i.e., features that start at position 4 or 5) that are among the top 10 or bottom 10 features, and 2) distal features (i.e., features that start at a position > 5) that are among the top 20 and bottom 20 features. Next, for each feature we selected the sequences for which the feature was > 0. Features for which no such sequences were found were removed from the analysis. If the number of corresponding sequences for a selected feature was larger than 10, we selected 10 of those sequences randomly, otherwise we considered all sequences for that feature. Finally, we made point mutations for the 1-bp features and double mutations for the 2-bp features selected as described above, and added both the wild-type and the mutant sequences to the microarray design. The results of this analysis are presented in Figure S4F, as squared Pearson correlation coefficients (R^2) between the predicted and the measured PBM log signal intensities over the probes corresponding to each selected feature. The high R^2 values (between 0.68 and 0.97 for Cbf1, and 0.64 and 0.92 for Tye7) show that our models are accurate at predicting the PBM log signal intensities. In addition to evaluating our predictions using the Pearson correlation coefficient, we also computed, for each selected feature, the fraction of mutated probes for which we correctly predicted whether the mutation will cause an increase or decrease in PBM signal intensity compared to wild-type sequence. As shown in Figure S4F, these numbers vary from 83% and 100% for features selected according to the Cbf1 model, and from 67% to 99% for features selected according to the Tye7 model. These validation results can be used in the future to improve the regression models by focusing on the sequences with the largest prediction errors.

To generate the "flank mutations" probes, we selected 5 genomic regions bound either by Cbf1 or Tye7 in vivo and generated all possible mutations of their 4-bp proximal flanks. The selected sequences are:

(1) CATGTCCGGCATCACGTGGTTATGATGTAG,
(2) ATTTGTCACAGTCACGTGATTACATTTAAG,
(3) ACGCGTTCGCTTCACGTGATGTTCCTTTTC,

(4) GTTGCTATATATCACGTGATCAATTTTTCC,
(5) CGTGTATACATTCACGTGACTATCTAGTAC.

Analyses of the "flank mutations" probes are presented in Figures S4C–S4E.

Similarly to the initial gcPBM, the "validation" microarray contains negative control probes selected so that they do not contain any putative Cbf1 or Tye7 binding sites. The "validation" array contains a total of 237 negative control sequences, and each sequence is present in 4 different replicate spots. All the other sequences on the "validation" array are present in 6 different replicate spots.

After running the PBM experiment on the "validation" array, we compared the PBM log intensity values for the wild-type sequences against the PBM log intensity values obtained for the same sequences in the initial gcPBM experiment, and we adjusted the PBM log intensity on the "validation" array to be in the same range as the values on the original array.

Statistical Analysis of the Differences in gcPBM Signal for 30-mer Probes Bound by Cbf1 versus Tye7 In Vivo

To assess whether differences in the in vitro binding preferences of Cbf1 and Tye7 contribute to differential DNA binding by these TFs in vivo, we examined whether the sequences preferred in vivo by each TF also have higher TF binding signal in vitro (Figure 6). As shown in Figure 6B, we first separated the gcPBM 30-mer probes based on in vivo specificity: blue for 30-mer probes selected from the 37 regions bound only by Cbf1 in vivo, and red for 30-mer probes selected from the 67 regions bound only by Tye7 in vivo (Figure 6B also shows, in gray, 30-mer probes selected from the 11 genomic regions bound by both Cbf1 and Tye7 in vivo; these probes were not used in the statistical analyses described here). Next, for each TF we filtered out the low-affinity gcPBM probes, defined as probes with log signal intensity less than two standard deviations above the mean signal observed for negative control probes. We eliminated these probes from the analysis because they are likely to be bound nonspecifically by the TFs. Using the specifically-bound gcPBM probes, we next compared the in vitro (gcPBM) signal for probes bound only by Cbf1 in vivo versus probes bound only by Tye7 in vivo (Figures 6B and 6C). We used the Kolmogorov-Smirnov test to determine whether the differences in in vitro binding signal between the two sets of probes (i.e., red versus blue points in Figures 6B and 6C) are statistically significant.

SUPPLEMENTAL REFERENCES

Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Peña-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T., et al. (2008). Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell *133*, 1266–1276.

Dudley, A.M., Aach, J., Steffen, M.A., and Church, G.M. (2002). Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. Proc. Natl. Acad. Sci. USA *99*, 7554–7559.

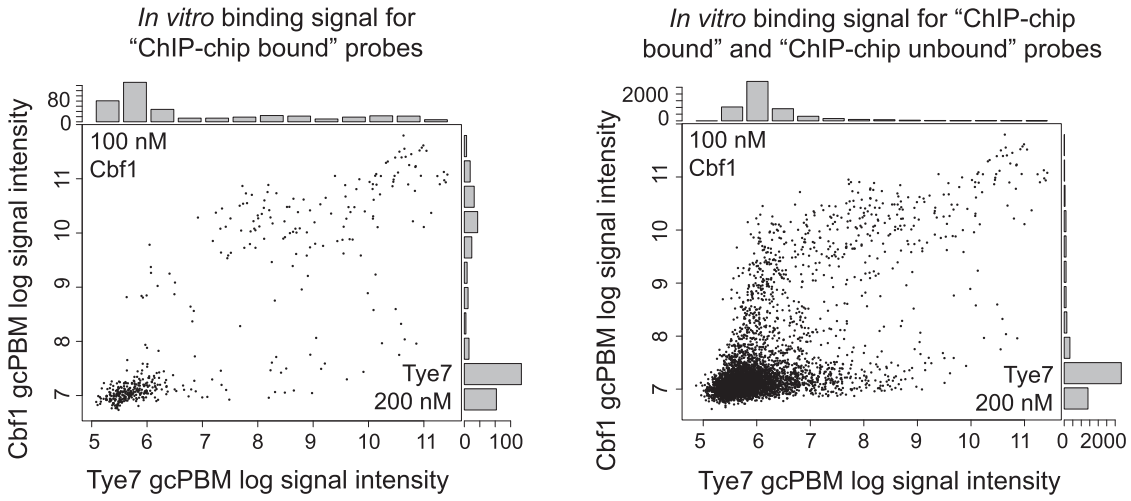


Figure S1. Comparison of In Vitro Binding of Cbf1 versus Tye7 to ChIP-chip Bound and ChIP-chip Unbound Probes on gcPBMs, Related to Figure 2

Concentrations of Cbf1or Tye7 in the PBM binding reactions were 100 nM and 200 nM, respectively. The plots show the natural logarithm of the normalized PBM fluorescence signal intensities, with higher numbers corresponding to higher affinity binding.

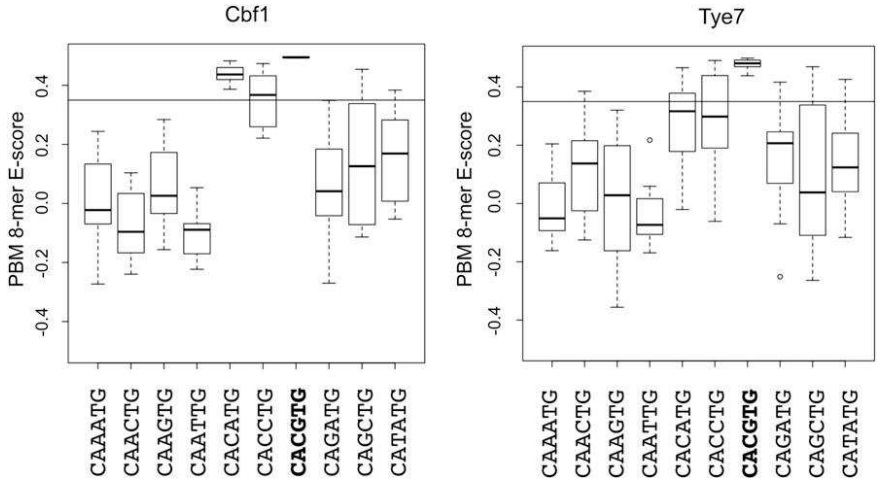


Figure S2. Universal PBM-Derived E-Box Specificities of Cbf1 and Tye7, Related to Figure 1

Box plots depict PBM E-scores (*y*-axis) for all 8-mers that contain a given E-box (*x*-axis). For each box plot, the central horizontal bar shows the median of the E-score distribution, the box's edges mark the 25th and 75th percentiles, and the whiskers represent the most extreme points of the distribution which were not determined to be outliers. For convenience, we include a horizontal bar at an E-score value of 0.35, above which we consider 8-mers to correspond to specific TF-DNA binding (Berger et al., 2006; Gordân et al., 2011).

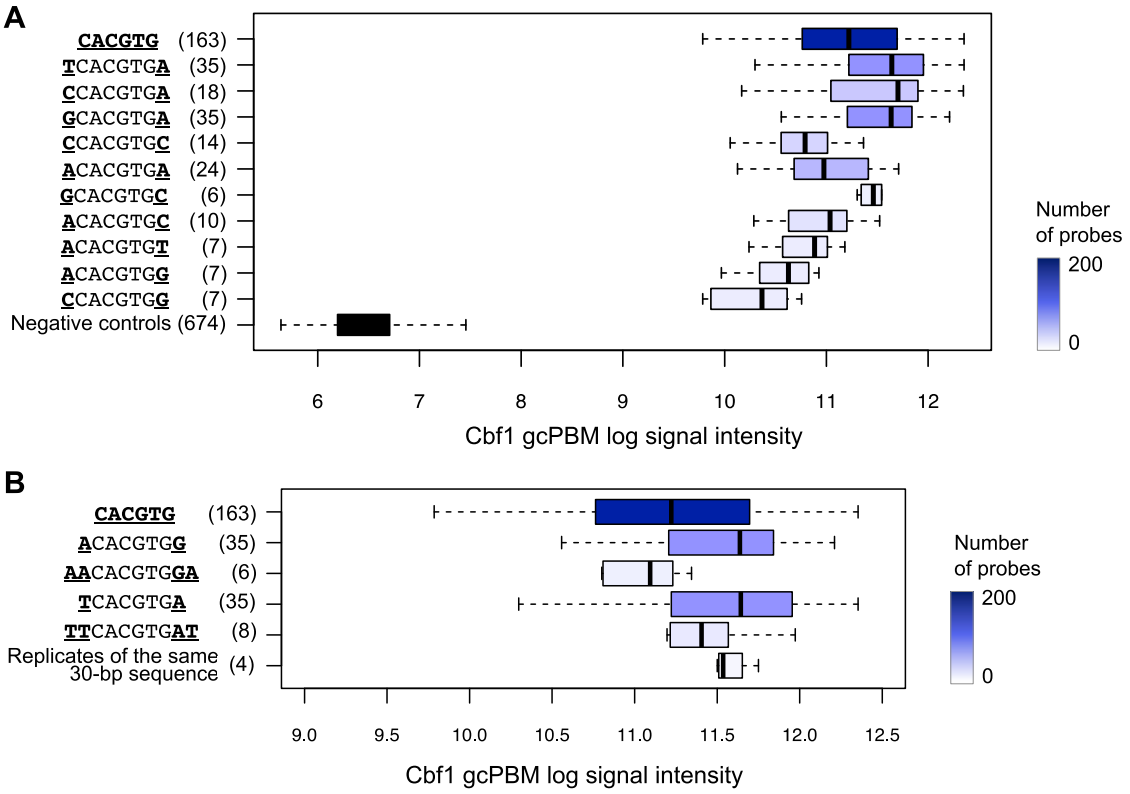
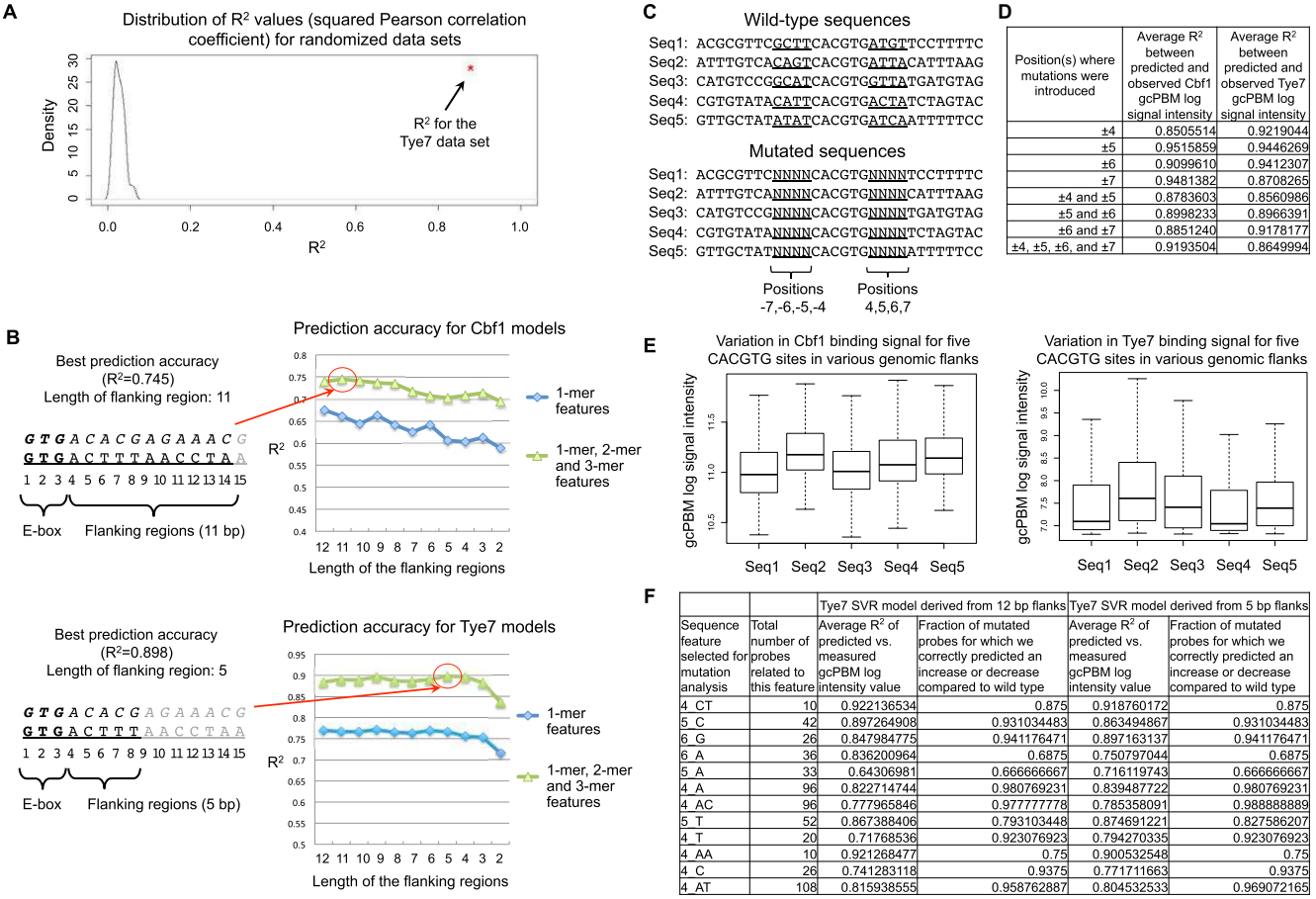


Figure S3. Variation in Cbf1-DNA Binding Signal due to Sequences Flanking the E-Box CACGTG, Related to Figure 3

(A) Variation in Cbf1-DNA binding signal for probes that contain the preferred E-box CACGTG (top box plot), or any of the possible 8-mers centered on this E-box (middle 10 box plots). The numbers in parentheses indicate how many probes contain each 6-mer or 8-mer, respectively (in either the forward or reverse complementary orientation). The negative control probes (bottom box plot) were chosen so that the maximum 8-mer E-score across the entire probe is < 0.3, *i.e.*, these probes do not contain any putative Cbf1 binding sites. Thus, the distribution of gcPBM log signal intensities for the negative control probes corresponds to nonspecific Cbf1-DNA binding.

(B) We observe a wide variation in DNA binding signal even when we restrict the analysis to probes containing specific 10-mers (*i.e.*, the E-box plus 5′ and 3′ proximal flanks). The box plot illustrates two examples of 10-mers showing a wide variation in Cbf1 binding signal. For comparison, we also show the variation among 4 replicate spots (bottom box plot) containing the same 30-bp sequence, which contains a TTCACGTGAT binding site (CCAGTGAAATTTCACGTGATGTAATCTGAT).

The boxes show the range between the 25th and 75th percentiles, the line within each box indicates the median, and the outer lines extend to 1.5 times the interquartile range from the box.



D

Position(s) where mutations were introduced	Average R^2 between predicted and observed Cbf1 gcPBM log signal intensity	Average R^2 between predicted and observed Tye7 gcPBM log signal intensity
±4	0.8505514	0.9219044
±5	0.9515859	0.9446269
±6	0.9099610	0.9412307
±7	0.9481382	0.8708265
±4 and ±5	0.8783603	0.8560986
±5 and ±6	0.8998233	0.8966391
±6 and ±7	0.8851240	0.9178177
±4, ±5, ±6, and ±7	0.9193504	0.8649994

F

		Tye7 SVR model derived from 12 bp flanks		Tye7 SVR model derived from 5 bp flanks	
Sequence feature selected for mutation analysis	Total number of probes related to this feature	Average R^2 of predicted vs. measured gcPBM log intensity value	Fraction of mutated probes for which we correctly predicted an increase or decrease compared to wild type	Average R^2 of predicted vs. measured gcPBM log intensity value	Fraction of mutated probes for which we correctly predicted an increase or decrease compared to wild type
4_CT	10	0.922136534	0.875	0.918760172	0.875
5_C	42	0.897264908	0.931034483	0.863494867	0.931034483
6_G	26	0.847984775	0.941176471	0.897163137	0.941176471
6_A	36	0.836200964	0.6875	0.750797044	0.6875
5_A	33	0.64306981	0.666666667	0.716119743	0.666666667
4_A	96	0.822714744	0.980769231	0.839487722	0.980769231
4_AC	96	0.777965846	0.977777778	0.785358091	0.988888889
5_T	52	0.867388406	0.793103448	0.874691221	0.827586207
4_T	20	0.71768536	0.923076923	0.794270335	0.923076923
4_AA	10	0.921268477	0.75	0.900532548	0.75
4_C	26	0.741283118	0.9375	0.771711663	0.9375
4_AT	108	0.815938555	0.958762887	0.804532533	0.969072165

Figure S4. Performance of SVR Models, Related to Figure 4

(A) Distribution of R^2 values (squared Pearson correlation coefficient) for 100 randomized data sets derived from the Tye7 data set. The parameters for the support vector regression were optimized using the same parameter search as for the real data set. The (*) indicates the R^2 obtained for the real data set.

(B) Accuracy of SVR models trained on flanking regions of different lengths. The best accuracy was obtained when 11-bp flanks were used to predict Cbf1 binding, and when 5-bp flanks were used to predict Tye7 binding.

(C) One category of probes on the validation microarray ("flank mutations" probes) was derived from 5 genomic regions containing CACGTG sites. We kept the E-box and the distal flanks constant, and generated all possible sequences in the 4-bp proximal flanks.

(D) Table shows Pearson correlation R^2 between the predicted and observed gcPBM log signal intensities for probes containing mutations at various positions relative to the E-box (e.g., to compute the R^2 for mutations at position 4 we selected those sequences that match one of the wild-type sequences at all positions except position 4, with the mutation occurring in either of the two flanks). Predictions were made using SVR models derived for Cbf1 and Tye7 from the 12-bp flanking regions (see main text for details).

(E) Box plots show the variation in gcPBM log signal intensity for Cbf1 and Tye7 across the mutated sequences generated from each of the genomic sequences in (C). Each box corresponds to one of the 5 genomic regions. The wide range of gcPBM log signal intensities obtained for each sequence shows the influence of the 4-bp proximal flanks on DNA binding. The variation we observe among the medians of the five gcPBM log signal intensity distributions (corresponding to the 5 sequences) shows the influence of the distal flanks. The differences between the Cbf1 and Tye7 boxplots show that the two TFs interact differently with these DNA sequences. The boxes show the range between the 25th and 75th percentiles, the line within each box indicates the median, and the outer lines extend to 1.5 times the interquartile range from the box.

(F) Performance of the Cbf1 and Tye7 SVR models on genomic sequences containing mutations at various positions in the flanking regions.

Cell PRESS
Open ACCESS

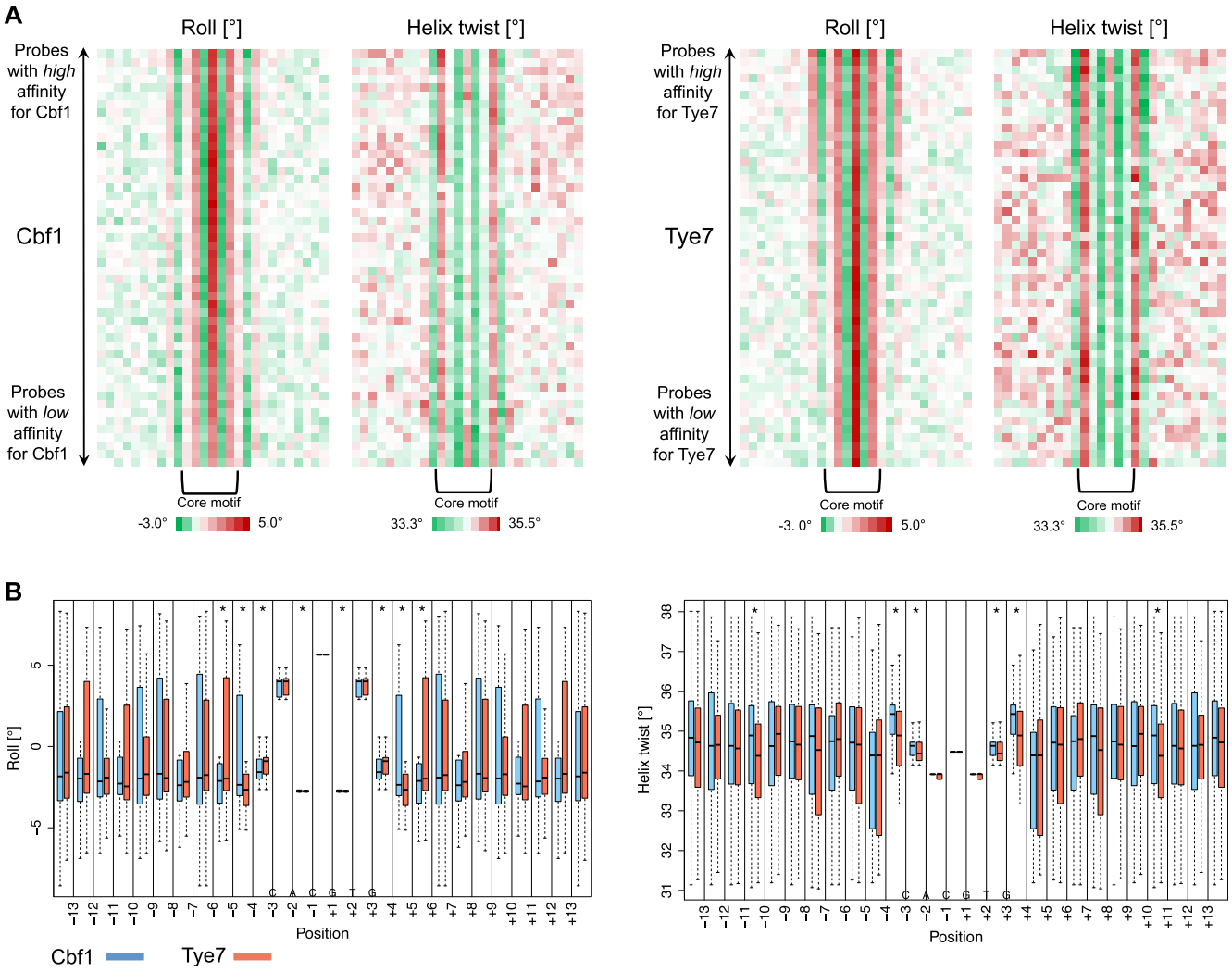


Figure S5. DNA Shape Analysis, Related to Figure 5

(A) Heat maps show the average roll and helix twist for the DNA sequences on the gcPBM. The sequences were sorted in decreasing order of the PBM signal intensity for either Cbf1 (left) or Tye7 (right), and grouped into 50 bins. Average DNA shape parameters were computed over the sequences in each bin.

(B) Box plots show the DNA shape variation due to flanks selected by Cbf1 (light blue) or Tye7 (light red). We compared the DNA shape parameters at individual positions between two groups of sequences that share the CACGTG E-box but are bound with higher affinity by either Cbf1 or Tye7 (see text for details). Asterisks (*) mark the positions at which we observed a significant difference (Mann-Whitney U $p < 0.05$) in roll or helix twist between the sequences preferred by the two TFs. The E-box is located at positions −3 to +3. The symmetry of the box plots is due to the shape predictions performed for both DNA strands, thus combining the left and right flanks.

The boxes show the range between the 25th and 75th percentiles, the line within each box indicates the median, and the outer lines show the range between the 5th and 95th percentiles.

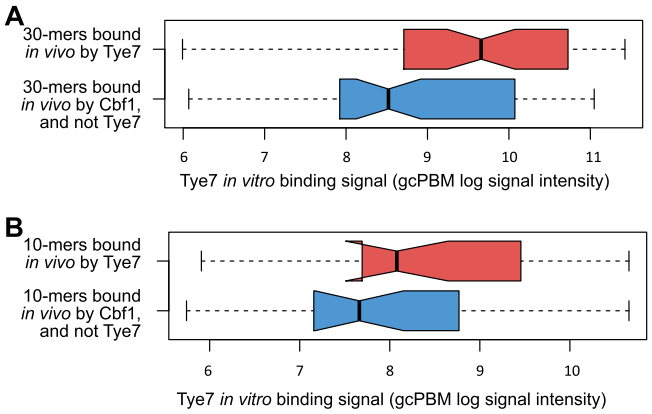


Figure S6. In Vitro Binding Differences in the Sequence Preferences of Cbf1 and Tye7 Partially Explain Differential In Vivo DNA Binding, Related to Figure 6

(A) Tye7 in vitro binding signal (gcPBM log signal intensity) for CACGTG-containing 30-mer probes that occur in Tye7_YPD (red), or in Cbf1_YPD but not in Tye7_YPD (blue). The difference in Tye7 PBM log signal intensity between the two sets of 30-mer probes is statistically significant (Kolmogorov-Smirnov (KS) $p = 0.0003612$), with probes bound by Tye7 in vivo having an overall higher DNA binding signal in vitro. We observed a similar trend for Cbf1, with probes bound by Cbf1 in vivo having a higher Cbf1 PBM signal than probes not bound in vivo by Cbf1 ($p = 0.04$; not shown).

(B) As in (A), except that CACGTG-containing 10-mers were tested within constant 10-bp DNA flanks on the gcPBM. The difference in Tye7 PBM log signal intensity between the two sets of 10-mers is not statistically significant (KS $p = 0.1696$).

The boxes show the range between the 25th and 75th percentiles, the line within each box indicates the median, and the outer lines extend to 1.5 times the interquartile range from the box.